LETTERS

A distal enhancer and an ultraconserved exon are derived from a novel retroposon

Gill Bejerano[1], Craig B. Lowe[1], Nadav Ahituv[2,3], Bryan King[1,4], Adam Siepel[1]†, Sofie R. Salama[1,4], Edward M. Rubin[2,3], W. James Kent[1] & David Haussler[1,4]

Hundreds of highly conserved distal *cis*-regulatory elements have been characterized so far in vertebrate genomes[1]. Many thousands more are predicted on the basis of comparative genomics[2,3]. However, in stark contrast to the genes that they regulate, in invertebrates virtually none of these regions can be traced by using sequence similarity, leaving their evolutionary origins obscure. Here we show that a class of conserved, primarily non-coding regions in tetrapods originated from a previously unknown short interspersed repetitive element (SINE) retroposon family that was active in the Sarcopterygii (lobe-finned fishes and terrestrial vertebrates) in the Silurian period at least 410 million years ago (ref. 4), and seems to be recently active in the 'living fossil' Indonesian coelacanth, *Latimeria menadoensis*. Using a mouse enhancer assay we show that one copy, 0.5 million bases from the neuro-developmental gene *ISL1*, is an enhancer that recapitulates multiple aspects of *Isl1* expression patterns. Several other copies represent new, possibly regulatory, alternatively spliced exons in the middle of pre-existing Sarcopterygian genes. One of these, a more than 200-base-pair ultraconserved region[5], 100% identical in mammals, and 80% identical to the coelacanth SINE, contains a 31-amino-acid-residue alternatively spliced exon of the messenger RNA processing gene *PCBP2* (ref. 6). These add to a growing list of examples[7] in which relics of transposable elements have acquired a function that serves their host, a process termed 'exaptation'[8], and provide an origin for at least some of the many highly conserved vertebrate-specific genomic sequences.

One of the most evolutionarily constrained regions in mammalian genomes is the ultraconserved element uc.338 (ref. 5), a mammal-specific 223-base-pair (bp) region perfectly conserved between human, mouse and rat, overlapping a short protein-coding exon of *PCBP2* (ref. 6). This small region was observed to have multiple paralogues within the human genome, overlapping protein-coding exons of otherwise unrelated genes, as well as conserved intronic and intergenic regions[9] (Supplementary Fig. S1). This region also has multiple homologues in coelacanth that are closer in sequence to the human ultraconserved element than many of its human paralogues (Fig. 1c).

Further scrutiny of the 1 million bases (Mb) of sequence available from the Indonesian coelacanth reveals that the match is contained in a 481-bp genomic repeat. A total of 59 closely related copies are found in all four different coelacanth genomic regions sequenced so far (Supplementary Table S1). Using these copies we reconstructed a consensus coelacanth sequence of this repeat (Supplementary Fig. S2). Despite the huge evolutionary separation between the two species, the coelacanth repeat consensus is 80% identical over 360 bp to the human region containing the ultraconserved element. We could find no significant similarity between the coelacanth sequence and any known repeat. However, SINE families are often generated by the fortuitous retroposition of a transfer RNA sequence into a location where it can provide a DNA polymerase III (Pol III) promoter for a retrotranscriptionally capable transcript[10,11]. Indeed, the coelacanth 5′ end is similar to the vertebrate serine tRNA, conserving the A and B boxes of an internal Pol III promoter, the 3′ end has a clear poly(A) region, and the sequence is free of internal oligothymidylate tracts (Fig. 1a). This, combined with the high copy number (extrapolated to about 10^5 copies genome-wide), low divergence between copies and evidence of target site duplications, indicates that the *L. menadoensis* sequences define a recently active SINE family, which we term the LF-SINE, for lobe-finned fishes (or 'living fossil') SINE. This family shares a weak 65-bp signature with two superfamilies of known SINEs (Supplementary Information S1).

Diverged LF-SINE copies are found in all available tetrapod genome drafts (Supplementary Table S3), as well as among the partial genomic data from a related coelacanth species and from multiple amniotes (Supplementary Table S4). We cannot detect significant LF-SINE matches in lungfish DNA sequence currently available (under 300 kilobases (kb)) or in genome drafts of available ray-finned fish or invertebrates (Supplementary Table S2). Nor is there any sequence similarity evidence in the public repositories indicating possible non-hereditary (horizontal) transfer from any other DNA source (Supplementary Information S2). It therefore seems that this SINE family was generated by a tRNA retroposition in a species of the ancestral Sarcopterygii and is specific to this clade (Fig. 2).

There are only several hundred recognizable copies of the LF-SINE in tetrapods for which we have genome drafts: 245 in human, 235 in dog, 394 in opossum, 699 in chicken and 26 in frog (Supplementary Table S3). These copies form orthology groups, in which each orthologue is in the same relative location with respect to the surrounding genes in all tetrapods where it is present (Fig. 1c). Each such group represents a single LF-SINE retroposition that occurred before the common ancestor of the species in the group. As expected from a recently active SINE, none of the 59 coelacanth instances have a human orthologue. However, multiple instances in tetrapods, including 29 in human (Supplementary Fig. S15), match the entire span of the reconstructed coelacanth SINE, including portions of its poly(A) tail, providing direct evidence for retroposition activity in the tetrapod lineage. This analysis establishes that virtually all retroposition events that generated mammalian LF-SINE instances predate the divergence of placental mammals and marsupials and that at least several of them, possibly all, predate the divergence of amniotes and amphibians. Examination of orthology groups using a very conservative test indicates that most human instances and their orthologues have evolved significantly more

[1]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. [2]DOE Joint Genome Institute, Walnut Creek, California 94598, USA. [3]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [4]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA. †Present address: Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA.

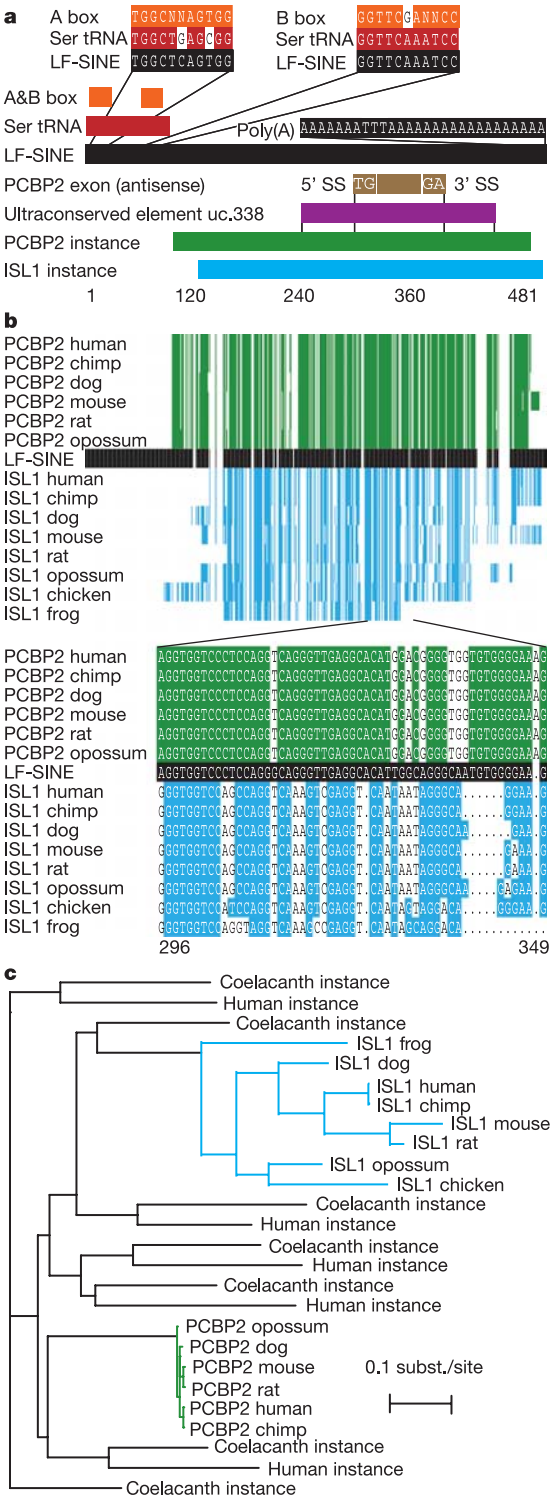

Figure 1 | Coelacanth SINE, human ultraconserved *PCBP2* exon and *ISL1* proximal enhancer share a common origin. **a**, Anatomy of the LF-SINE and its relation to an exapted tetrapodal distal enhancer near *ISL1*, and the ultraconserved exon of *PCBP2*, exonized from the reverse strand. SS, splice site. **b**, Alignment of multiple species instances of the *PCBP2* exonized element, and *ISL1* proximal LF-SINE enhancer, with the reconstructed coelacanth SINE. Filled squares (matches) and white spaces (tetrapodal inserts) are with respect to the coelacanth sequence. **c**, A maximum-likelihood joint phylogeny of selected LF-SINE instances from multiple species. The orthologous copies are shown to form monophyletic subtrees, whereas the additional instances serve to demonstrate the remarkable overall similarity between human and coelacanth instances.

slowly than would be expected assuming neutrality (Supplementary Information S3). This indicates that most detectable instances of the LF-SINE in tetrapods might have been exapted into cellular roles benefiting the host, subjecting them to purifying selection. In some cases the exapted tetrapod instance is remarkably close to the coelacanth SINE, indicating that the active LF-SINE in coelacanth might have changed very little over more than 410 Myr of independent evolution. The dispersion of coelacanth instances over many subclades in the evolutionary tree for these elements (Fig. 1c) precludes the possibility of recent horizontal transfer from tetrapods to coelacanth.

Most human instances of LF-SINEs are either intergenic (163 of 245; 66%; 107 more than 100 kb from a known gene) or intronic (68; 28%), and a smaller subset (14; 6%) overlap documented exons. We cannot find transcriptional evidence or predictions indicating that the human LF-SINEs are active as small RNAs or are involved in antisense regulatory transcripts. However, LF-SINE instances are found preferentially near genes involved in transcriptional regulation and neuronal development, indicating possible exaptation to form distal *cis*-regulatory regions (Supplementary Information S4).

To test this hypothesis, we picked a likely enhancer candidate and tested it *in vivo* using mouse transient transgenics. The *ISL1* gene encodes a LIM homeobox transcription factor that is required for motor neuron differentiation[12] and is expressed in motor and sensory neurons during vertebrate embryogenesis[13]. An *ISL1* proximal LF-SINE instance, significantly conserved between mammals, chicken and frog, lies 488 kb downstream of *ISL1*, in a 1.4-Mb gene desert that is home to two confirmed distal enhancers[13](Fig. 3a). The relative ordering and proximity to *ISL1* of the previously characterized enhancers and the LF-SINE instance represent an ancient organization that is invariant in frog, chicken, opossum, mouse and human (Supplementary Fig. S8).

The human *ISL1* proximal LF-SINE instance was cloned upstream of a mouse minimal heat shock 68 (*Hsp68*) promoter coupled to the β-galactosidase (*lacZ*) reporter gene and injected into the pronuclei

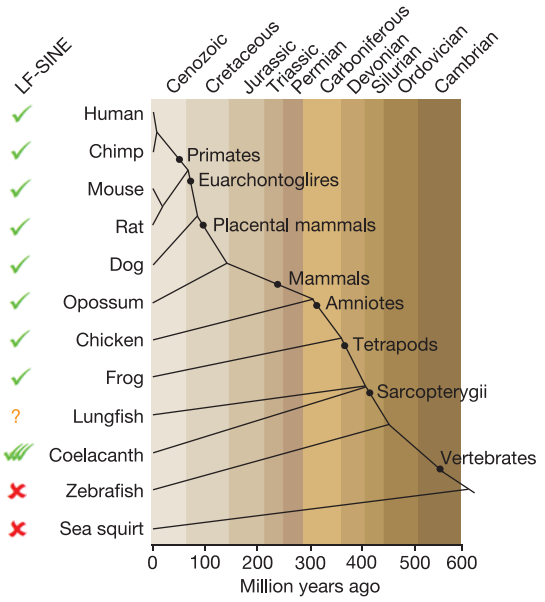

Figure 2 | Phylogeny of chordate genomes searched for instances of the LF-SINE. LF-SINE copies were found in the draft genomes of all terrestrial vertebrates shown and in genomic regions available from two coelacanth species. The LF-SINE was not found in very partial genomic data from lungfish, nor in any available draft genome of non-sarcopterygian vertebrates and invertebrates, including the two shown here. Temporal estimates are taken from ref. 4 and later sources. One tick, 25–700 copies in genome draft; three ticks, 59 copies in 1 Mb of DNA; question mark, no copies in less than 300 kb of DNA; cross, no copies in genome draft.

of fertilized mouse oocytes. The resulting embryos were analysed at embryonic day 11.5 (E11.5) by whole-embryo staining for *lacZ* activity (see Methods). Eight of nine independent *ISL1* proximal LF-SINE transgenic embryos showed consistent expression in the head and spinal cord region, the dorsal apical ectodermal ridge and genital eminence; in addition, four of nine embryos showed staining in the trigeminal ganglion (Fig. 3). Horizontal sections demonstrate specific colocalization of the *ISL1* proximal LF-SINE-driven *lacZ* reporter and murine *Isl1* RNA in neural tissues (Fig. 4). These expression patterns clearly recapitulate aspects of *Isl1* expression in developing motor neurons at this developmental stage[13,14]. The novel and the two previously described enhancers in this region drive a very similar pattern of reporter gene expression at E11.5. They may drive expression distinctively at a different time point, perhaps later in development, as data for the two known enhancers seem to indicate[13]. Our combined functional and evolutionary analysis indicates that this LF-SINE instance might have been exapted as an *ISL1* enhancer before the divergence of the tetrapods and still functions in this capacity today. This constitutes a proof that mobile elements give birth to distal enhancers.

The *ISL1* proximal LF-SINE instance and the instance overlapping ultraconserved region uc.338 have conserved a very similar portion of the ancient LF-SINE (Fig. 1). However, one serves as a distal enhancer, and the other as an alternatively spliced exon. To gain a better understanding of exonization, we examined all 19 LF-SINE instances that were exapted into protein-coding mRNAs (Supplementary Table S6). The affected proteins, encoded by *PCBP2*, *SMARCA4*, *EEF1B2*, *TCERG1*, *PTDSR*, *RORA*, *GRID1*, *ATF2*, *FLJ22833*, *ARHGAP6*, *KIAA1409*, *NT5C2*, *LRP1B*, *DHX30*, gg-*DMTF1*, gg-*PPP2R2C*, gg-*SHFM1*, xt-*MBNL1* and *JGI-49280*, are unrelated. Only a single pair of them shares a structural domain (helicase). All 19 derived exons are antisense to the original LF-SINE transcript. In 17 of 19 cases a new exon is formed in the middle of the coding region. Only canonical splice sites are used, similarly yet distinct from primate specific Alu-SINE exonization[15] (Supplementary Information S5). Exapted exons start in all three possible reading frames. Sixteen of 17 are alternatively spliced, potentially leaving the original functional isoforms intact while evolution optimized the function of the novel isoform[16]. Eleven of 17 introduce an early stop codon, predicted to trigger nonsense-mediated decay[17]. Often the most evolutionarily conserved regions are the LF-SINE-derived intronic regions immediately flanking the exons, indicating the possible presence of exapted regulatory elements. Taken together, these observations do not indicate a common protein structural modification induced by exonization of the LF-SINE. Rather, LF-SINE exaptation might be used to regulate the protein levels, including in *PCBP2*, in which the ultraconserved exon might be involved in cellular localization[18], dimerization[19] and post-transcriptional auto-regulation[20], as well as in *SMARCA4* (*BRG1*; ref. 21) and *LRP1B* (ref. 22; Supplementary Information S5).

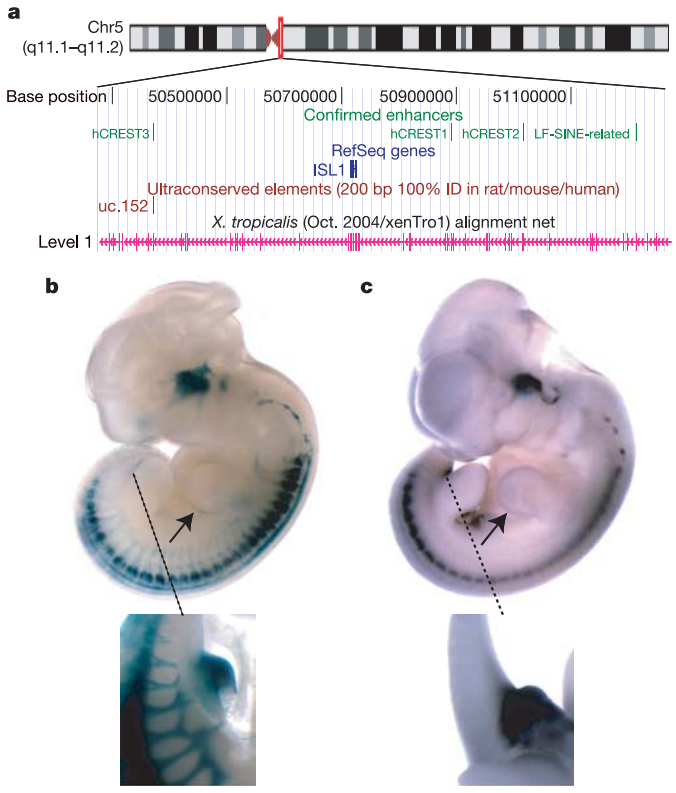

Figure 3 | A SINE-derived distal enhancer near *ISL1*. **a**, A 1-Mb pericentromeric neighbourhood of *ISL1* holds three previously confirmed enhancers[13] (hCREST1, hCREST2 and hCREST3 = uc.152), and the novel LF-SINE-derived enhancer, 488 kb downstream of *ISL1*. The genomic organization of *ISL1* and the four enhancers is conserved between human and frog (*Xenopus tropicalis*). **b**, Expression pattern of a representative reporter gene construct driven by the human *ISL1* proximal LF-SINE in a transient transgenic mouse at E11.5. **c**, This pattern recapitulates major aspects of the expression pattern of the mouse *Isl1* gene at E11.5, assayed with whole-mount *in situ* hybridization. Enlargements show the genital eminence and arrows indicate the staining of the dorsal apical ectodermal ridge.

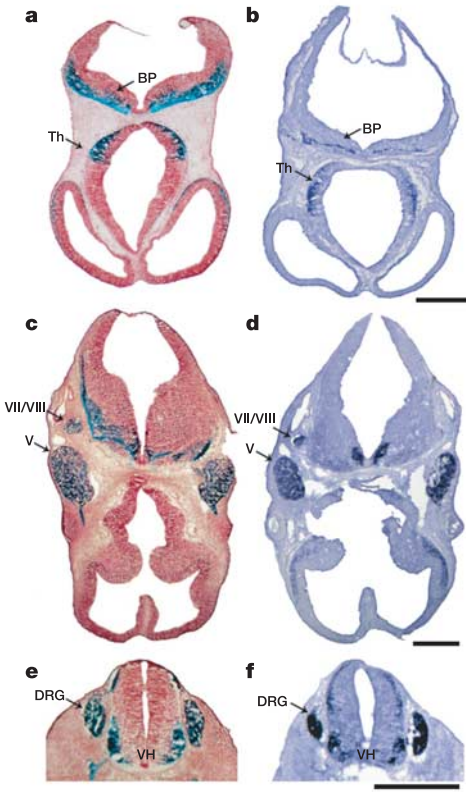

Figure 4 | Neural-specific expression driven by *ISL1*-proximal-LF-SINE recapitulates *Isl1* expression. Horizontal sections through E11.5 mice. **a**, **c**, **e**, *LacZ* staining in blue from the *ISL1*-LF-SINE-*LacZ* transient transgenic embryos with a neutral red counterstain. **b**, **d**, **f**, *In situ* RNA hybridization of *Isl1* in wild-type embryos. Matched level sections show corresponding expression patterns in the developing thalamus (Th) and basal plate (BP) in the brain (**a**, **b**), the trigeminal (V) ganglion and facio-acoustic (VII/VIII) ganglia in the head region (**c**, **d**), and the dorsal root ganglion (DRG) and the lateral region of the ventral horn (VH) of the spinal cord (**e**, **f**; thoracic sections). In **a**–**d** posterior is up; in **e** and **f** dorsal is up. Scale bars, 0.5 mm.

After discovering mobile DNA elements, Barbara McClintock suggested that they were fundamentally involved in gene regulation[23], an idea further developed by Britten and Davidson, who speculated on the benefit of obtaining similar control regions for a 'battery' of co-regulated genes through exaptation[24]. At least 50% of our genome originates from characterized transposon-derived DNA[25]. Although the early systematic theories of their role in gene regulation were not confirmed, it seems possible that, because these elements optimize their interaction with the host machinery under strong, virus-like evolutionary pressures, they are a particularly fecund source of evolutionary innovations, including new gene regulatory elements, and these are at times exapted by the host to improve its own fitness[7]. If so, it is possible that many more of the one million conserved vertebrate genomic elements originated from ancient retroposon families. In support of this hypothesis we find thousands of paralogue families of highly conserved non-coding sequences in the human genome[9,26], as well as individual exons with multiple non-coding paralogues (for example, Supplementary Fig. S12), of unknown origins (Supplementary Information S6).

The SINE families that are active in the eight tetrapods for which we have so far obtained draft genomes have all been restricted to specific clades, indicating rather recent origins and thus a rather rapid turnover of active SINE families on an evolutionary timescale. The Indonesian coelacanth may be different in that its LF-SINEs, and independently discovered Deu-SINEs (H. Nishihara, A. Smit and N. Okada, personal communication), have apparently remained active for more than 400 Myr with very little change. By preserving what in other species would be transient transposon families, the coelacanth acts, in a sense, as a living molecular fossil. The remaining 99.9% of its genome, as yet unsequenced[27], may very well hold precious traces of additional events that helped shape our own evolution.

METHODS

Enhancer analysis. The *ISL1* proximal LF-SINE instance was amplified from human DNA (BD Bio- sciences Clontech) by polymerase chain reaction (PCR) with the following primers: forward, 5′-AACATCTTGAAAAGAAGATCT AAGC-3′; reverse, 5′-AAGCTGCTTTTAAAACTGTATCTTC-3′. The amplified DNA was cloned into the *Hsp68-lacZ* vector[28]. The construct was then purified and injected into pronuclei as described previously[29]. Embryos were harvested at E11.5, and transgenic embryos were identified by PCR of *lacZ*, using DNA from yolk sac and the following *lacZ* primers: forward, 5′-TTTCCATGTTGCC ACTCGC-3′; reverse, 5′- AACGGCTTGCCGTTCAGCA-3′. Expression of *lacZ* was assayed in all embryos, with 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-Gal; Sigma), as described previously[29]. See also ⟨http://enhancer.lbl.gov/aboutproject.html⟩. The *ISL1* proximal LF-SINE transient transgenics were further analysed by examining 20-μm horizontal sections made from cryo-preserved embryos. Sections were counterstained with neutral red (Sigma; 0.3% w/v in PBS) to reveal the tissues. For *in situ* RNA hybridizations a murine *Isl1*-containing plasmid (IMAGE no. 3419119) was used to make sense and antisense digoxigenin-labelled RNA probes. *In situ* hybridizations on wild-type E11.5 embryos were performed as described previously[30]. Stained sections were photographed with a PowerShot G6 digital camera (Canon) mounted on a dissecting microscope.

Computational analysis. The computational methods are described in Supplementary Information.

Received 27 November 2005; accepted 2 March 2006.
Published online 16 April 2006.

1. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7 (2005).
2. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
3. International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
4. Graur, D. & Li, W.-H. *Fundamentals of Molecular Evolution* 2nd edn, Appendix I (Sinauer, Sunderland, Massachusetts, 2000).
5. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
6. Makeyev, A. V. & Liebhaber, S. A. The poly(C)-binding proteins: A multiplicity of functions and a search for mechanisms. *RNA* **8,** 265–278 (2002).
7. Brosius, J. The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* **118,** 99–116 (2003).
8. Gould, S. & Vrba, E. Exaptation—a missing term in the science of form. *Paleobiology* **8,** 4–15 (1982).
9. Bejerano, G., Haussler, D. & Blanchette, M. Into the heart of darkness: Large-scale clustering of human non-coding DNA. *Bioinformatics* **20,** i40–i48 (2004).
10. Weiner, A. M. SINEs and LINEs: The art of biting the hand that feeds you. *Curr. Opin. Cell Biol.* **14,** 343–350 (2002).
11. Deininger, P. L. & Batzer, M. A. Mammalian retroelements. *Genome Res.* **12,** 1455–1465 (2002).
12. Pfaff, S. L., Mendelsohn, M., Stewart, C. L., Edlund, T. & Jessell, T. M. Requirement for LIM homeobox gene *Isl1* in motor neuron generation reveals a motor neuron-dependent step in interneuron differentiation. *Cell* **84,** 309–320 (1996).
13. Uemura, O. *et al.* Comparative functional genomics revealed conservation and diversification of three enhancers of the *isl1* gene for motor and sensory neuron-specific expression. *Dev. Biol.* **278,** 587–606 (2005).
14. Caton, A. *et al.* The branchial arches and HGF are growth-promoting and chemoattractant for cranial motor axons. *Development* **127,** 1751–1766 (2000).
15. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* **300,** 1288–1291 (2003).
16. Makalowski, W. Genomics. Not junk after all. *Science* **300,** 1246–1247 (2003).
17. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA* **100,** 189–192 (2003).
18. Chkheidze, A. N. & Liebhaber, S. A. A novel set of nuclear localization signals determine distributions of the alphaCP RNA-binding proteins. *Mol. Cell. Biol.* **23,** 8405–8415 (2003).
19. Kim, J. H., Hahm, B., Kim, Y. K., Choi, M. & Jang, S. K. Protein–protein interaction among hnRNPs shuttling between nucleus and cytoplasm. *J. Mol. Biol.* **298,** 395–405 (2000).
20. Waggoner, S. A. & Liebhaber, S. A. Identification of mRNAs associated with αCP2-containing RNP complexes. *Mol. Cell. Biol.* **23,** 7055–7067 (2003).
21. Gunduz, E. *et al.* Genetic and epigenetic alterations of BRG1 promote oral cancer development. *Int. J. Oncol.* **26,** 201–210 (2005).
22. Li, Y., Lu, W. & Bu, G. Striking differences of LDL receptor-related protein 1B expression in mouse and human. *Biochem. Biophys. Res. Commun.* **333,** 868–873 (2005).
23. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA* **36,** 344–355 (1950).
24. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* **46,** 111–138 (1971).
25. Kazazian, H. H. Jr. Mobile elements: Drivers of genome evolution. *Science* **303,** 1626–1632 (2004).
26. McEwen, G. K. *et al.* Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Res.* doi:10.1101/gr.4143406 (2006).
27. Danke, J. *et al.* Genome resource for the Indonesian coelacanth, *Latimeria menadoensis*. *J. Exp. Zoolog. A* **301,** 228–234 (2004).
28. Kothary, R. *et al.* A transgene containing *lacZ* inserted into the dystonia locus is expressed in neural tube. *Nature* **335,** 435–437 (1988).
29. Poulin, F. *et al. In vivo* characterization of a vertebrate ultraconserved enhancer. *Genomics* **85,** 774–781 (2005).
30. Feldheim, D. A. *et al.* Topographic guidance labels in a sensory projection to the forebrain. *Neuron* **21,** 1303–1313 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the various sequencing consortia and research groups for the numerous genomic regions used in this study; A. Hinrichs, M. Diekhans, R. Harte, G. Barber and the UCSC browser team, M. Shoukry, I. Plajzer-Frick, S. Chanan and V. Afzal for technical help; R. Baertsch, T. Furey, E. Margulies and J. S. Pedersen for sharing unpublished data; and W. Miller, M. Blanchette, D. Feldheim, M. Nobrega, M. Ares, C. Sugnet, M. Dermitzakis and J. Brosius for discussions. Research conducted at University of California Santa Cruz was supported by the National Human Genome Research Institute and the Howard Hughes Medical Institute; Research conducted at the E.O. Lawrence Berkeley National Laboratory was supported by grants from the Programs for Genomic Application, the NHLBI and performed under a Department of Energy Contract, University of California.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.B. (jill@soe.ucsc.edu).